Assignment

Evaluation:

Modern will evaluate each entry in two phases:

- Assessment of the code submitted, coding standards, best practices, formatting, module/class size, etc
- Evaluation of the presentation (of a small subset of entries, reasonably good on the above criteria)

Submission:

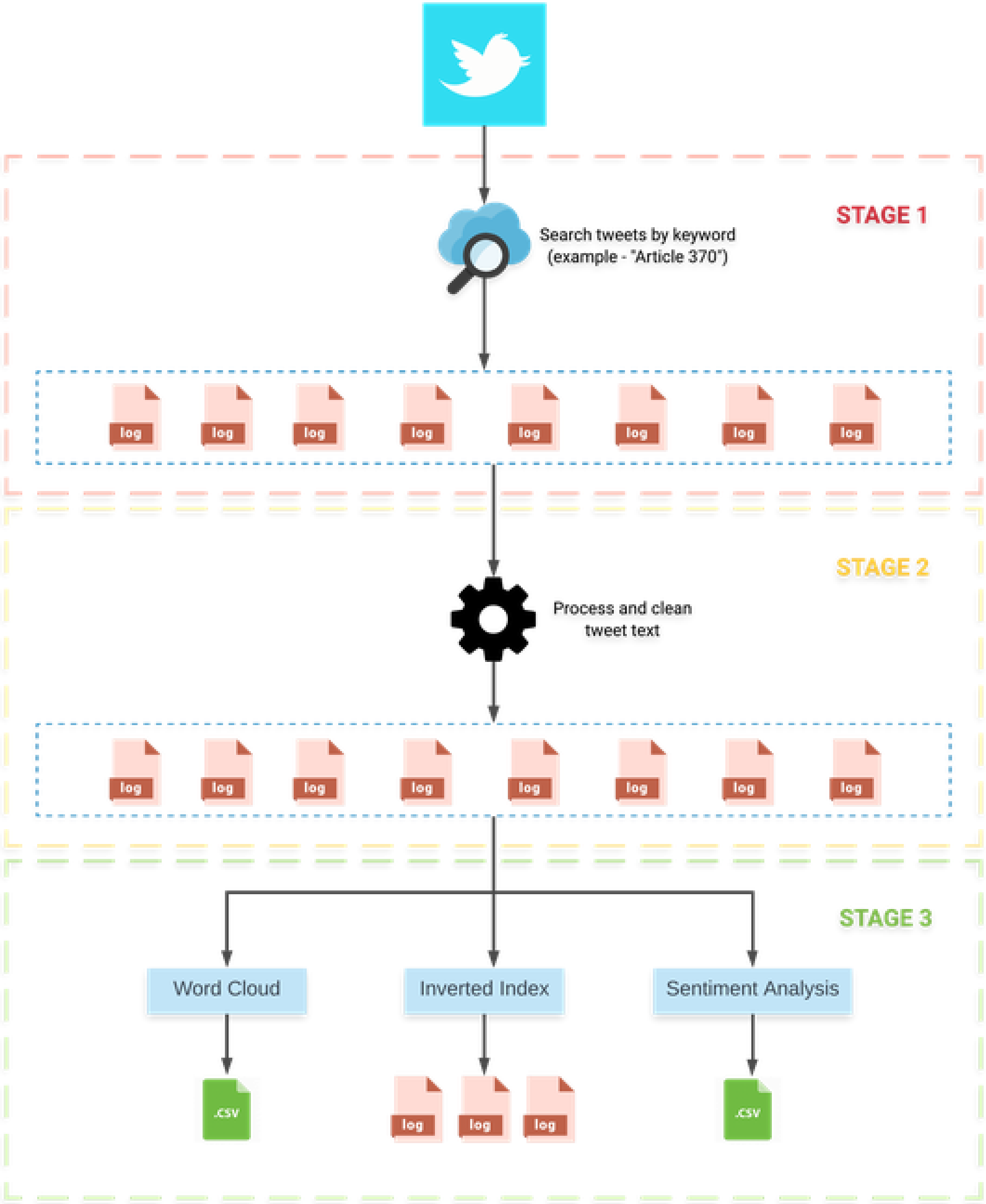
- Write proper tests and documentation.
- Build application strictly on Python 3 or Java 8
- Deliver your source code strictly in a public Github repo.
- At the end of the run for each stage/subproblems, the command should produce metrics like time taken, the number of records written /read, etc.

Problem Statement:

Create an application that finds tweets using the keyword (for example - '***covid19***'), stores raw data in log files, cleanses the raw data, and performs specific text analysis.

There are three stages to the problem, with individual scores for each stage; for an entry to be eligible for the evaluation, it has to provide a solution to Stagea 1 at least.

- Stage 1 - Find recent tweets using the keyword and store the raw data in log files.
- Stage 2 - Process and clean the raw data and save the processed data in log files
- Stage 3 - Perform the following tasks on processed data
  - Problem A - Generate a word cloud
  - Problem B - Create an inverted index for hashtags, mentions, and words appearing in the tweets
  - Problem C - Perform Sentiment Analysis and publish distribution by place



✓ Scoring Methodology for Phase 1 evaluation

- Coding Standards (10 points)
- Stage 1 (20 points)
- Stage 2 (20 points)
- Stage 3
    - Problem A (5 points)
    - Problem B (15 points)
    - Problem C (20 points)

Stage 1

You will write a program that takes a keyword as input and finds recent tweets (100,000 tweets) containing that keyword using twitter's official API and stores the raw data (tweet JSON object) in log files ensuring that the log file should not exceed 10000 records or 5 Mb in size (whichever constraint is met first). You will have multiple log files for the raw data; choose the naming convention for these files wisely, as the program will reuse files in Stage 2. You can also use a third-party library (Tweepy for python).

Example of Tweet JSON

```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "@eric @kelvin 1\/ Today we\u2019re sharing our vision for
the future of
           the Twitter API platform!\nhttps:\/\/t.co\/XweGngmxlP
#future #platform",
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter Platform news,
updates &
                    events. Need technical help? Visit
                    https:\/\/twittercommunity.com\/ \u2328\ufe0f
#TapIntoTwitter"
  },
  "place":
{
  "attributes":{},
   "bounding_box":
  {
     "coordinates":
     [[
           [-77.119759,38.791645],
           [-76.909393,38.791645],
           [-76.909393,38.995548],
           [-77.119759,38.995548]
     ]],
     "type":"Polygon"
  },
   "country":"United States",
   "country_code":"US",
   "full_name":"Washington, DC",
   "id":"01fbe706f872cb32",
   "name":"Washington",
   "place_type":"city",
   "url":"http://api.twitter.com/1/geo/id/0172cb32.json"
},
  "entities": {
    "hashtags": [
    ],
      "urls": [
        {
          "url": "https:\/\/t.co\/XweGngmxlP",
          "unwound": {
            "url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\
/3xo1c",
            "title": "Building the Future of the Twitter API Platform"
          }
        }
      ],
      "user_mentions": []
    }
  }
```

- Your program should prompt the user to input the keyword.
- You will write a tweet JSON object per LINE in the log file. The program can then parse each line individually.
- After successfully storing data in log files, your program should print the following statistics in the console after termination.
    - Time is taken to run the program.
    - Average size (in bytes) of a tweet JSON
    - Number of log files created to store data
    - The average number of records in a log file
    - Average size (in KB) of a log file

Stage 2

Process and clean the raw data stored in Stage 1 and keep the processed data in new log files, ensuring that the log file should not exceed 10000 records or 5 Mb in size (whichever constraint is met first). You will have multiple log files for the processed data; choose the naming convention for these files wisely, as the program will reuse them in Stage 3. It would be best to create a new JSON object to store only relevant information in log files.

Structure of new JSON object

```
  {
    "tweetId": "850006245121695744",
    "rawTweet": "@eric @kelvin one \/ Today we\u2019re sharing our vision
for the future of the
                Twitter API platform!\nhttps:\/\/t.co\/XweGngmxlP
#future #platform",
    "tidyTweet": "Today sharing vision future Twitter API platform future
platform",
    "tokenizedTweet":
['today','sharing','vision','future','twitter','API','platform',
                      'future','platform'],
    "listMentions": ['eric','kelvin'],
    "listHashtags": ['future','platform'],
    "countRetweets": 3,
    "placeName": 'Washington'
  }
```

tidyTweet - Clean the tweet text for each record in the raw data

1. Remove URLs
2. Remove Mentions
3. Remove Reserved words (RT, FAV)
4. Remove Emojis
5. Remove Smileys
6. Remove Stop Words - A stop word is a commonly used word (such as "the", "a", "an", "in").
7. Remove Punctuation and Special Characters
8. Apply Stemming and Lemmatization techniques for text normalization (**Not Mandatory**)

tokenizedTweet - list of words appearing in the tidyTweet

listMentions - list of mentions in the rawTweet

listHashtags - list of hashtags in the rawTweet

countRetweets - number of retweets of the given tweetId

placeName - the name of the place associated with the given tweetId

- You will write a new JSON object per LINE in the log file. The program can then parse each line individually.
- After successfully storing data in log files, your program should print the following statistics in the console after termination.
  - Time is taken to run the program.
  - Average size (in bytes) of a new JSON
  - Minimum and maximum size of new JSON objects
  - Number of log files created to store data
  - The average number of records in a log file
  - Average size (in KB) of a log file
  - Count of URLs removed.

Stage 3

Use the processed data of Stage 2 and write a program to find solutions to the following subproblems.

Problem A - Generate Word Cloud

Find the count of each word in the entire data set of tweet texts and save the result in CSV format in descending order of count.

Example

```
words, count
vaccine, 97653
death, 95763
comorbid, 85475
```

- You have to save the output in CSV format.
- After successfully storing data in a CSV file, your program should print the following statistics in the console after termination.
  - Time is taken to run the program.
  - Top 5 most frequent words in the data set
  - Number of unique words in the data set
  - Size (in KB) of the CSV file

Problem B - Create an Inverted Index

Write a program that produces an inverted index that gives, for every hashtag, the list of tweet ids it appears in and saves the data in the log file. It would be best to create a new JSON object to store only relevant information in the log file. Also, enable index search so a user can search for a hashtag and the list of tweet ids.

You must create an Inverted index and store it in a different log file for each of the following.

- Hashtags
- Mentions
- Unique words in the data set

Structure of new JSON object

```
{
  "#foreignpolicy": ['850006245121695744','557076155121895534']
}
```

- Provide a function to search words in the created indexes
- You will write a **new** JSON object per LINE in the log file. The program can then parse each line individually.
- After successfully storing data in the log file, your program should print the following statistics in the console after termination.
    - Time is taken to run the program.
    - Top 5 most common hashtags in the data set
    - Number of unique hashtags in the data set
    - Size (in KB) of the log file

Problem C - Sentiment Analysis and its distribution by place

Write a program that tells you whether a tweet expresses a positive sentiment, negative sentiment, or neutral and publish the distribution of the result by place in a CSV file

Example

```
place,positive,negative,neutral
placeA, 70%, 20%,10%
placeB, 50%, 10%, 40%
```

- You have to save the output in CSV format.
- After successfully storing data in the log file, your program should print the following statistics in the console after termination.
    - Time is taken to run the program.
    - Count of places in favor of Article 370
    - Count of places against Article 370
    - Size (in KB) of the CSV file

Output Format:

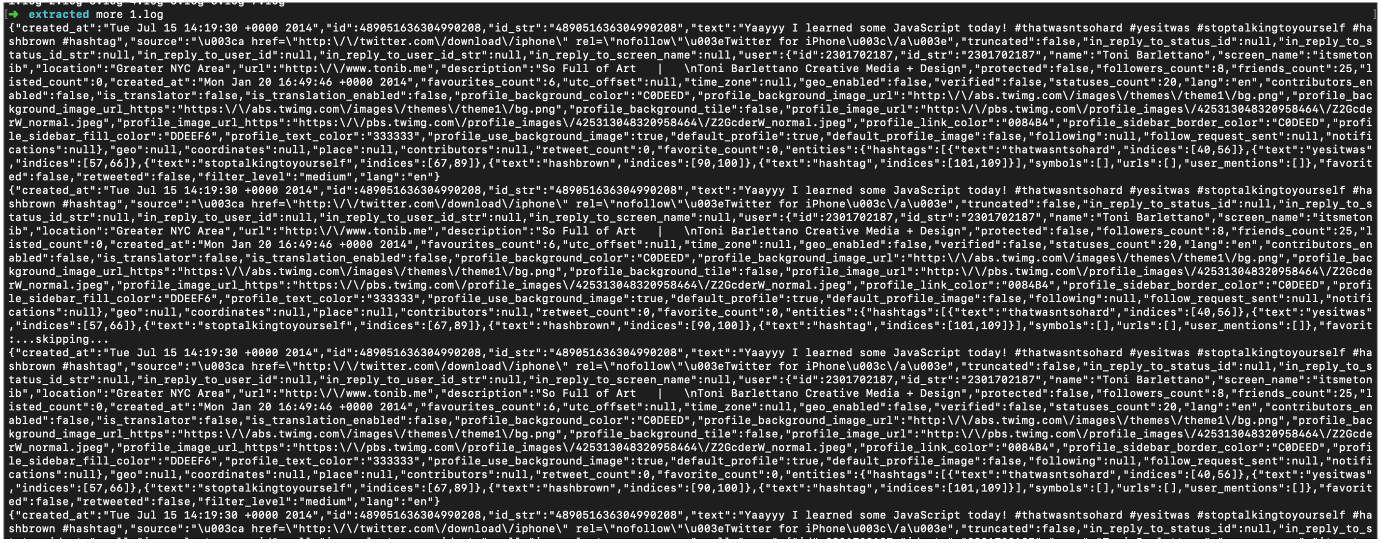
Folder structure

Save all the output files in the following structure. Create a root folder with the name as a current timestamp. Under it, create the 'moderndata' folder. Create individual folders for each of the stages.

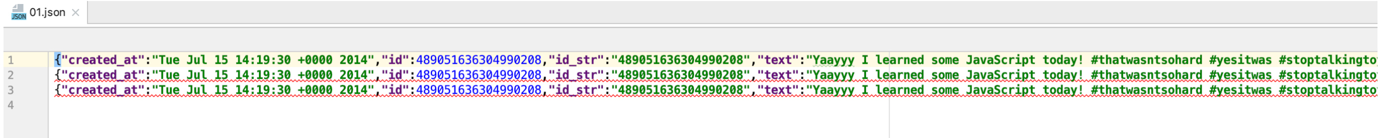
```
1565353790875
 moderndata
      cleansed
       1.log
       2.log
       3.log
       4.log
       5.log
      extracted
       1.log
       2.log
       3.log
 4.log

 5.log

 6.log

 7.log

wordcounts

 1.csv

inverted-index

 hashtag.log

 mentions.log

 wordcloud.log

sentiment

1.csv
```

Log file

You will write a tweet JSON object per LINE in the log file. The program can then parse each line individually.



```
[➜  extracted more 1.log
{"created_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingtoyourself #ha
shbrown #hashtag","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_s
tatus_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":2301702187,"id_str":"2301702187","name":"Toni Barlettano","screen_name":"itsmeton
ib","location":"Greater NYC Area","url":"http:\/\/www.tonib.me","description":"So Full of Art   |   \nToni Barlettano Creative Media + Design","protected":false,"followers_count":8,"friends_count":25,"l
isted_count":0,"created_at":"Mon Jan 20 16:49:46 +0000 2014","favourites_count":6,"utc_offset":null,"time_zone":null,"geo_enabled":false,"verified":false,"statuses_count":20,"lang":"en","contributors_en
abled":false,"is_translator":false,"is_translation_enabled":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_bac
kground_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/425313048320958464\/Z2Gcde
rW_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/425313048320958464\/Z2GcderW_normal.jpeg","profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profi
le_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifi
cations":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[{"text":"thatwasntsohard","indices":[40,56]},{"text":"yesitwas"
,"indices":[57,66]},{"text":"stoptalkingtoyourself","indices":[67,89]},{"text":"hashbrown","indices":[90,100]},{"text":"hashtag","indices":[101,109]}],"symbols":[],"urls":[],"user_mentions":[]},"favorit
ed":false,"retweeted":false,"filter_level":"medium","lang":"en"}
{"created_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingtoyourself #ha
shbrown #hashtag","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_s
tatus_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":2301702187,"id_str":"2301702187","name":"Toni Barlettano","screen_name":"itsmeton
ib","location":"Greater NYC Area","url":"http:\/\/www.tonib.me","description":"So Full of Art   |   \nToni Barlettano Creative Media + Design","protected":false,"followers_count":8,"friends_count":25,"l
isted_count":0,"created_at":"Mon Jan 20 16:49:46 +0000 2014","favourites_count":6,"utc_offset":null,"time_zone":null,"geo_enabled":false,"verified":false,"statuses_count":20,"lang":"en","contributors_en
abled":false,"is_translator":false,"is_translation_enabled":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_bac
kground_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/425313048320958464\/Z2Gcde
rW_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/425313048320958464\/Z2GcderW_normal.jpeg","profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profi
le_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifi
cations":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[{"text":"thatwasntsohard","indices":[40,56]},{"text":"yesitwas"
,"indices":[57,66]},{"text":"stoptalkingtoyourself","indices":[67,89]},{"text":"hashbrown","indices":[90,100]},{"text":"hashtag","indices":[101,109]}],"symbols":[],"urls":[],"user_mentions":[]},"favorit
:...skipping...
{"created_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingtoyourself #ha
shbrown #hashtag","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_s
tatus_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":2301702187,"id_str":"2301702187","name":"Toni Barlettano","screen_name":"itsmeton
ib","location":"Greater NYC Area","url":"http:\/\/www.tonib.me","description":"So Full of Art   |   \nToni Barlettano Creative Media + Design","protected":false,"followers_count":8,"friends_count":25,"l
isted_count":0,"created_at":"Mon Jan 20 16:49:46 +0000 2014","favourites_count":6,"utc_offset":null,"time_zone":null,"geo_enabled":false,"verified":false,"statuses_count":20,"lang":"en","contributors_en
abled":false,"is_translator":false,"is_translation_enabled":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_bac
kground_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/425313048320958464\/Z2Gcde
rW_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/425313048320958464\/Z2GcderW_normal.jpeg","profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profi
le_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifi
cations":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[{"text":"thatwasntsohard","indices":[40,56]},{"text":"yesitwas"
,"indices":[57,66]},{"text":"stoptalkingtoyourself","indices":[67,89]},{"text":"hashbrown","indices":[90,100]},{"text":"hashtag","indices":[101,109]}],"symbols":[],"urls":[],"user_mentions":[]},"favorit
ed":false,"retweeted":false,"filter_level":"medium","lang":"en"}
{"created_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingtoyourself #ha
shbrown #hashtag","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_s
```



```
{"created_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingto
{"created_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingto
{"created_at":"Tue Jul 15 14:19:30 +0000 2014","id":489051636304990208,"id_str":"489051636304990208","text":"Yaayyy I learned some JavaScript today! #thatwasntsohard #yesitwas #stoptalkingto
```

Bonus Questions

1. The difference in time taken by programs to run by use of Parallel Processing
2. Visual Word Cloud
3. Store processed data in SQLite and enable full-text search
4. Analyze the co-occurrence of words in a tweet